Self-Employment Income Reporting on Surveys

by

Christian Imboden
Bowling Green State University

John Voorheis
U.S. Census Bureau

Caroline Weber
University of Kentucky

CES 23-19 **April 2023**

Abstract

We examine the relation between administrative income data and survey reports for self-employed and wage-earning respondents from 2000 - 2015. The self-employed report 40 percent more wages and self-employment income in the survey than in tax administrative records; this estimate nets out differences between these two sources that are also shared by wage-earners. We provide evidence that differential reporting incentives are an important explanation of the larger self-employed gap by exploiting a well-known artifact – self-employed respondents exhibit substantial bunching at the first EITC kink in their administrative records. We do not observe the same behavior in their survey responses even after accounting for survey measurement concerns.

Keyword*:* Income Reporting, Survey Accuracy, Measurement Error, Tax Evasion, Tax Avoidance

JEL Classification: C83, H24, H26

* Imboden: cimbode@bgsu.edu; Voorheis: john.l.voorheis@census.gov; Weber: caroline.weber@uky.edu. The authors would like to thank David Agrawal, Chris Bollinger, Iuliia Shybalkina, Glen Waddell, and Jim Ziliak for helpful comments. We appreciate comments and feedback from participants at seminars at University of Kentucky, Association for Budgeting and Financial Management, and the National Tax Association Annual Conference. David Shi has provided exemplary research assistance.

1 Introduction

There is a long literature on the differences between the income reports individuals provide to survey and tax authorities (e.g. Bound and Krueger, 1991; Pischke, 1995; Bollinger, 1998; Bound et al., 2001; Bollinger et al., 2019). This literature highlights measurement error issues in the survey data, such as the "common man" hypothesis and survey non-response. Some recent work also considers measurement error in tax administrative data, including mismatch, and finds less evidence of the common-man hypothesis (Kapteyn and Ypma, 2007; Meijer et al., 2012; Abowd and Stinson, 2013; Bollinger et al., 2018). Taken as a whole, the literature suggests that differences in survey and administrative income reports are caused by complex interactions of multiple factors. These factors have explained meaningful, but relatively small, average gaps between survey and tax administrative records. One understudied and potentially substantial source of income reporting differences is the respondent's primary source of income. Self-employed respondents have a different set of incentives in the decision to accurately report their incomes to the tax authority, as compared with wage-earning respondents. They may also face different challenges in reporting their income accurately to surveys.

To examine the role of income source – wages or self-employment – in discrepancies (hereafter referred to as "gaps") between survey and administrative data, we link data from the Current Population Survey's Annual Social and Economic Supplement (ASEC) to administrative records from the Social Security Administration Detailed Earnings Records (DER), as well as information from the IRS Form 1040. We restrict our sample to those who report earnings to both the DER and ASEC, and therefore we do not address the issue of

survey item nonresponse.[1] Additionally, we largely abstract away from the issue of income miscategorization (Abraham et al., 2021) by utilizing income concepts that group multiple detailed categories of income together. Thus, our baseline earnings measure is wages + self-employment income. Data is described in detail in Section 2.

We find greater divergence between survey and administrative earnings reports for self-employed taxpayers, relative to wage-earners. Specifically, we find that self-employed taxpayers report 50 percent more to the ASEC than to the IRS, on average. This difference is about 40 percentage points larger than the same difference for wage-earners. This gap has remained stable over our sample period. It is relatively larger for those with above-median survey income. It is not being driven by inherent survey reporting differences by these two types of taxpayers for all income types: the gap disappears for interest and dividend income, which has the same informational reporting requirements for both wage-earners and the self-employed.

Survey measurement errors will not explain the larger gap for self-employed taxpayers (even in part) unless survey measurement errors are exacerbated for self-employment income in a way that, on average, substantially overstates self-employment income on the survey. We conduct numerous robustness checks in tables and figures along these lines. We confirm that rounding, income type mismatch, and the timing of the IRS filings and survey responses, and other factors, do not play a substantial role in the size of the self-employed gap relative to the wage-earner gap.

We consider tax reporting incentives as a possible causal mechanism for the relative gaps. The self-employed and wage-earners face substantially different incentives to accurately

[1] See (e.g. Bollinger et al., 2019) for an in depth analysis of survey nonresponse.

report income to tax administrative authorities. Wages are subject to employer information reporting and withholding, whereas self-employment income is not. Random IRS audits suggest only 37% of income subject to little or no informational reporting (such as income from sole proprietorships) is reported to the IRS, while 99 percent of wages are reported to the IRS (Slemrod and Bakija, 2008, p. 257). To the extent that evasion or avoidance behaviors are not reflected in survey reports, this will lead to a larger divergence of survey and administrative income for self-employed taxpayers.[2]

We find a large mass of the self-employed whose reported incomes for tax purposes bunch near the first kink of the Earned Income Tax Credit (EITC) schedule,[3] but there is no prominent bunching in survey income, nor in either data set for wage-earners. Prior research shows that bunching at the first EITC kink among the self-employed is largely driven by evasion decisions (Chetty et al., 2013). To the extent that wage-earners are subject to greater information reporting and withholding requirements, there are plausibly fewer evasion opportunities for wage-earners.

To provide causal evidence that differential reporting incentives drive the larger gap among the self-employed, we test the following hypothesis: one and two dependent self-employed taxpayers have similar income levels and report similar amounts to the survey; in contrast, two-dependent taxpayers report more to the tax authority than do one dependent taxpayers because the first EITC kink is higher for two-dependent taxpayers. This is exactly what we find. The bunching at the first EITC kink moves, depending on the number of

[2]We discuss tax evasion and avoidance in this paper exclusively to better understand what surveys measure when they ask questions about earnings. This research cannot identify any individual case of tax avoidance or evasion.

[3]The first EITC kink occurs at the income level where the phase-in of the EITC ends. This amount increases with the number of dependents.

dependents, but there is no such shift in the survey data – one and two dependent self-employed taxpayers have no sizable differences in the distribution of their survey income. As a result of this pattern, we find that two-dependent taxpayers have a 21 percentage point smaller gap relative to one dependent taxpayers (relative to any baseline gaps based on number of dependents in the wage-earner sample). This gap is approximately unchanged when we control for survey measurement concerns and demographic factors.

Our work contributes to several literatures beyond the survey measurement literature already cited. First, we contribute to the literature on the discrepancies between survey and administrative data for self-employed income (Bee et al., 2021; Bee and Rothbaum, 2019). Our work examines the types of measurement error issues that are specific to self-employed taxpayers, compares their measurement error issues with wage-earners, and provides causal evidence of a mechanism behind these differences. Second, we contribute to the literature on self-employed households (e.g. Bruce, 2000; Carroll et al., 2000; Gentry and Hubbard, 2000; Gale and Brown, 2013) and their evasion/avoidance decisions across the income distribution (DeBacker et al., 2020). We provide evidence in Appendix A that is consistent with Hurst et al. (2014), who argue that there are some remnants of tax evasion/avoidance in survey reports. We then make an additional contribution – by their same logic, not all evasion/avoidance in administrative records appears to be incorporated into survey records. Finally, we contribute to the literature on tax knowledge and tax rate perceptions (e.g. Chetty et al., 2013; Gideon, 2017; Ballard and Gupta, 2018): taxpayers that bunch precisely at the first EITC kink for tax purposes often do not report this decision on the survey.

2 Data

We link microdata from survey and two administrative records sources covering tax years 2000 through 2015 (survey years 2001 through 2016). Survey data come from the Census' internal use version of the Current Population Survey Annual Social Economic Survey (ASEC) and tax administrative data come from Social Security Administration Detailed Earnings Records (DER), with some supplemental items from IRS Form 1040 tax returns.

The monthly CPS survey primarily collects employment information from about 60,000 American households, and is the source of the Bureau of Labor Statistics' published unemployment rates. One individual per address responds on behalf of all individuals at that address. The ASEC is a supplement to the CPS, conducted annually around each March, and collects more detailed household income information about the previous year. It is the source of the Census Bureau's published official poverty rate.[4] Households may be surveyed in the ASEC for one or two consecutive years. Households that appear in two subsequent years are called the overlap sample.

The ASEC collects individual information on a variety of income categories, including wages, self-employment earnings, interest and dividends, child support, etc. Some ASEC data are imputed, and higher incomes are top coded. Relative to the public use version of the survey, the internal use version used here has higher income top codes. The ASEC also collects demographic information such as age, gender, household size, marital status, educational attainment, hours worked, industry, and home-ownership.

Our primary source of tax administrative data are the Detailed Earnings Records (DER),

[4]Most households in sample in March are given the ASEC supplement, but some households in sample in either February or April are also given the ASEC in order to oversample Hispanic and low-income households.

which are collected by the Social Security Administration based on IRS Form 1040 and W-2 data. These detailed records contain all wage earnings (reported on form W-2) and self-employed earnings subject to Social Security and Medicare taxes for all individuals in the Social Security system.[5] Since the self-employment income in the DER is sourced from Schedule SE—which is intended to calculate self-employment taxes on Form 1040—the DER contains only positive (which is to say, taxable) self-employment income amounts. The Census Bureau receives an extract of the DER for all linkable individuals in the ASEC.

Additionally, we link supplemental information directly from IRS Form 1040. Forms 1040 contain detailed income information for all taxable and some non-taxable income categories, and the U.S. Census Bureau receives some of the Form 1040 fields.[6] We utilize information on filing status, number of dependents, number of dependents claimed on the EITC, filing week, and indicators for filing schedules C (sole proprietorship income), D (capital gains/losses), E (partnership, LLC, S corporation, rent and royalty income) and F (farm income).

The US Census Bureau's data linkage infrastructure allows for the linkage of ASEC and DER/IRS records using anonymous, unique, time-invariant identifiers called Protected Identity Keys (PIKs). PIKs are assigned using the Person Identification Validation System (PVS), which is a probabilistic matching algorithm that uses personally identifiable information (PII) to link individuals to a reference file Layne et al. (2014). This reference file is a modified version of the SSA Numident. The Census Numident is the universe of individuals who have received Social Security Numbers (SSN), and contains PII including the SSN itself,

[5]This includes all positive non-farm and farm self-employment income from Form 1040 Schedules C & F.

[6]These include wages and salaries, dividends received, both taxable and non-taxable interest income, social security income, rental and royalty income, adjusted gross income, and total money income. Total money income is our most comprehensive measure of income.

as well as age, date of birth, sex, race and address.[7] We match all ASEC respondents from survey years 2001 through 2016 with DER and IRS Form 1040 data for the previous tax year. We construct two indicators based on the quality of the matches to use as robustness checks in Table 1. One indicator captures only the very best match in our sample – a match for someone in the taxpaying unit based on SSN. Census stopped collecting SSNs in tax year 2004. The second indicator includes SSN matches as well as the next best match – a match based on name, date of birth, and zip code.

We drop any person records for whom we are not able to obtain a PIK and match across all three data sources (ASEC, DER, and IRS). We aggregate the ASEC and DER data to the tax-unit level when we match it with the IRS data. Moreover, we drop any observations with missing, imputed, or truncated values. We also drop any data points where no wages + self-employment income was reported either to the DER or to the ASEC. This last restriction focuses our paper exclusively on the intensive margin. Because of these restrictions, our sample is not necessarily representative of the U.S. population as a whole, nor is it representative of the U.S. taxpaying population. Parallel work at the Census Bureau (Bee et al., 2023) is investigating ways to weight linked survey and administrative records to recover nationally representative blended income estimates, efforts which will inform future work in this area.

We define taxpayers as "self-employed" if 75 percent or more of their DER wages + self-employment income comes from self-employment. All other individuals are "wage-earners." Because we aggregate up to the taxpaying unit, we often refer to each agggregated obser-

[7]There are multiple vintages of the Numident reference file, each of which has the best available PII information for a given individual.

vation as a taxpayer as a shorthand.[8] Within taxpaying units, we determine the primary earner and create indicator variables for the primary earner's age, gender, hours worked, education level (less than high school diploma completed, high school diploma completed with no college, some college completed with no degree, and bachelors degree completed), industry, and race/ethnicity (Black, Native American, Asian, Hispanic, White). We restrict the sample to those with a primary earner aged 20 - 65 who are not dependents, to focus on the working-age population. We also create additional survey measurement indicators for our analysis. One measure indicates whether or not any reported survey income is rounded to the nearest $1,000. Another measure indicates whether the Form 1040 was published in the data bank (which generally corresponds to the filing week) in the first 10 weeks of the year as a rough proxy for whether the taxpayer filed their Form 1040 with the IRS before completing the ASEC survey.

To abstract away from the previously mentioned issue of income miscategorization on surveys (i.e. reporting wage income as self-employment income), our baseline measure of earnings is wages + self-employment income in both the survey and administrative data (this includes farm income).[9] We also construct a measure of total money income from the survey data, which can be compared to the measure of total money income available in the Census Bureau's 1040 extracts. [10] Amounts are in real 2010 dollars using the PCE deflator.

In addition to income miscategorization in survey data, it is possible that there is income

[8]Our findings are not sensitive to this aggregation decision – we demonstrate this in Table 1 Column (5) – and it allows us to examine differential reporting incentives as a causal mechanism for the larger gap among the self-employed.

[9]While we generally abstract away from this miscategorization issue, we do examine more elaborate miscategorization concerns in Section 4.

[10]Total money income combines wages, self-employment income (including farm income), interest and dividends, rental income, unemployment compensation, social security income, alimony, retirement income, and other income.

miscategorization in administrative data. There is prior evidence that some individuals have self-employment income, but do not report that income in a SE tax filing, instead reporting it as wages or other income (Collins et al., 2019). This form of miscategorization will not affect the calculation of "gaps" when using our broader total money income measure. However, it will make these taxpayers more likely to be categorized as wage-earners, which may create attenuation bias when comparing the gaps of the self-employed with the gaps of wage-earners.

3 Empirical Methods

Our analysis utilizes kernel densities and regression analysis. We compare the self-employed sample to the wage-earner sample. This nets out all forms of measurement error that are consistent across these two types of earners and allows us to follow the more recent literature that allows for measurement error in both survey and administrative sources (Kapteyn and Ypma, 2007; Abowd and Stinson, 2013; Meijer et al., 2012; Bollinger et al., 2018). Of course, it may be that certain measurement error issues are magnified among the self-employed sample; we consider these as possible explanations for our findings as well.

We present two types of kernel densities in this paper. The first type of kernel density plots survey and administrative earnings distributions separately for the self-employed and wage-earners. The initial figures plot income in real (2010) dollars. We also present figures using a renormalization of income. To renormalize income, we adjust all taxpayers' incomes proportionally so that their first EITC kink occurs at the same income level for all taxpayers regardless of marital status, number of dependents, or tax year. To do this, we choose one

dependent unmarried taxpayers as our base unit of analysis.[11] This renormalization will aid in the analysis of reporting incentives as a mechanism for self-employment mismeasurement.

We will sometimes compare the densities of self-employed taxpayers across two groups (e.g. round income and not round income). To allow the comparison to focus only on discrepancies driven by the grouping and not underlying income or demographic differences, we will reweight one group to be more like the other using inverse probability weighting (IPW). We obtain weights by regressing a group indicator on $1,000 ASEC income bins and a set of demographic variables.[12]

Our second type of kernel density exploits our linked survey and tax administrative data by plotting the "gaps" between administrative and survey reported incomes. These gaps are defined as:

$$Gap_{it} = \ln(Admin_{it}) - \ln(Survey_{it})$$

where $Admin_{it}$ and $Survey_{it}$ represent a taxpayer i in year t's incomes as reported to the DER and ASEC, respectively. Generally, income is measured as wages + self-employment income, but we consider two alternative measures of income in Table 1: interest and dividend income and total money income. We plot kernel densities of these gaps separately for wage-earners and the self-employed.

[11] To construct renormalized income, we use one dependent unmarried taxpayers as a baseline group. In real 2010 dollars, the first EITC kink is always $8,970 for these taxpayers during the period of analysis. Then, all incomes are multiplied by the ratio of the value of the first EITC kink point for the one dependent unmarried group to the first EITC kink for each taxpayer.

[12] These include age, age squared, and indicator variables for female, hours worked last week (under 10 hours, 60 hours and over, and every 5 hour increment in between), number of dependents (0, 1, 2 or more), education (less than high school, high school, some college, bachelor's degree), married, race and ethnicity (White, Black, Asian, Native American, and Hispanic), and two-digit industry code.

Our regressions take the following form:

$$Gap_{it} = \beta_0 + \beta_1 SE_{it} + \beta_2 SE_{it} \cdot X_{it} + \beta_3 X_{it} + u_{it}, \tag{1}$$

where Gap_{it} is the gap defined above for taxpayer i in year t. SE_{it} is our measure of self-employment – we define taxpayers as self-employed if over 75 percent of DER wage and self-employment income comes from self-employment sources.[13]. X_{it} are covariates of interest or relevant control variables that vary across regression specifications (survey income bins, year fixed effects, demographics and survey accuracy), and u_{it} is a residual. To keep the tables concise, we report only the coefficients on SE and interaction terms of interest. All standard errors are clustered by taxpayer (taxpayers in the overlap sample are observed in the data set twice).

4 Results

We begin our analysis in the top panel of Figure 1 by plotting kernel densities of positive survey and administrative wage and self-employment income separately for wage-earner and self-employed taxpayers whose survey and administrative income are both less than \$100,000.[14] There are almost 510,000 tax-unit observations in this figure and 5.2 percent of them are self-employed. The "gap" between self-employed survey and administrative data is

[13]Appendix Table B.1 considers alternative cut points; the gaps are substantial regardless of the cut point chosen and largest for the highest cut points. Appendix Table B.1 also considers defining self-employment based on survey income instead; we expect this is potentially a noisier measure of true self-employment status given the miscategorization issues documented in the literature (Abraham et al., 2021). Not surprisingly, the estimates of the self-employment gap are smaller based on this measure, but the gap remains sizeable.

[14]We require both survey and administrative income to be below \$100,000 so that the same taxpayers appear in both figures. This restriction does not meaningfully affect the figure. The \$100,000 cut is imposed because the density thins out substantially above \$100,000.

substantial – survey income is $13,000 (62%) higher than administrative income, on average. The gap is much smaller for wage-earners. The known bunching at the first EITC kink in the administrative data, primarily driven by tax evasion (Chetty et al., 2013), seems apparent in the administrative data in this figure; however, it is imprecise because EITC kink points vary depending on family composition.

To explore behavior around the first EITC kink more directly, we modify the kernel densities by renormalizing taxpayers' earnings so that the first EITC kink aligns for all taxpayers in the bottom panel of Figure 1. The first EITC kink is marked with a vertical black dashed line. We observe that the peak of the administrative self-employed distribution occurs at the first EITC kink. There is no similar peak in the survey data for self-employed taxpayers (or for either source among wage-earners). Hence, one explanation of the divergence between survey and administrative reports is the differential reporting incentives that self-employed taxpayers face; they have clear incentives and opportunities to engage in tax avoidance or evasion in the administrative data, and no such incentives to do so when reporting to the survey. We return to a more careful examination of this story as a causal mechanism for the larger gap between survey and administrative reports for the self-employed in Section 4.1.

In Figure 2, we plot a kernel density of the wage and self-employment income gap between administrative and survey reports directly for the self-employed (black line) and wage-earners (gray dashed line). This figure includes the same taxpayers as the top panel of Figure 1. For wage-earners, there is a large spike at zero (representing a large mass of taxpayers who report similar incomes to both sources) and the average gap is relatively small (-6.8 percent); in contrast, the peak of the density is negative for the self-employed and there is a long, fat left tail (the mean gap is -50.0 percent). The average gap is 43.2 percentage points larger in

absolute value (=-50+6.8) for self-employed taxpayers than for wage-earners. The median gap is similar; it is 41.3 percentage points larger in absolute value (=-0.435+0.022) for self-employed taxpayers than for wage-earners. The median gap is an order of magnitude larger for the self-employed than for wage-earners. The average gap is also documented in the first column of Table 1; the estimated gap reported there is similar, but it is estimated on a slightly different sample which no longer excludes those above $100,000 (this is for visual reasons only). Our baseline estimate in Table 1 Column (1) finds that the gap between administrative and survey data is 40 percentage points larger for the self-employed than it is for wage-earners.

Table 1 Column (2) adds year fixed effects and $1,000 income bins in survey wage + self-employment income as a robustness check.[15] This absorbs variation across years and survey income bins; to the extent that self-employed status and the gap itself both vary across these dimensions, our baseline estimate may have been biased. These estimates are even larger than our baseline estimates – the gap increases in absolute value to 48.6 percentage points.

The top panel of Figure 3 examines whether this gap has changed over time. We interact SE in equation (1) with tax year indicators and plot the estimated self-employed gap (relative to the wage-earner gap) separately for each tax year. The reference year is 2007. We overlay a second set of estimates that excludes those with survey income less than $500 as very small amounts of income can generate extreme outliers.[16] While a few of these yearly estimates are statistically different from zero, they are all small relative to the baseline estimated gap and there is no trend in these estimates over time.

[15] We exclude these fixed effects in our baseline estimates so that the constant represents the gap between survey and administrative income for wage-earners.

[16] While these extreme outliers are occasionally observable in individual years, the overall estimates are not affected by these.

The bottom panel of Figure 3 examines how this gap varies by income. We interact SE in equation (1) with an indicator for having above median survey income. We separately repeat this exercise using an indicator based on median administrative income[17] We report the total gap for the self-employed (relative to wage-earners) for each of these groups in the figure. This figure indicates that the gap increases with survey income – it is significantly larger for above-median income taxpayers, and this difference is statistically significant. The difference is the opposite for the administrative income split; those with below median administrative income have a sizable gap, but this gap shrinks to less than 10 percentage points for those above median administrative income. These patterns can potentially be explained if the administrative data reflect more evasion/avoidance than survey data because of differential reporting incentives for the self-employed – a possibility we provide evidence on in Section 4.1. DeBacker et al. (2020) observe similar distinctions when plotting evasion amounts among audited taxpayers across pre- and post-audit income (the pre-audit data is the same as our administrative data, and the post-audit data potentially more like our survey data, with at least some evasion eliminated).

Figures 1 and 3 and the first columns of Table 1 highlight much larger discrepancies between survey and administrative wage and self-employment income for the self-employed relative to wage-earners. These are mostly time-invariant and increase with survey income. Alone, this result has important implications for research measuring self-employed earnings or poverty. We return to these implications in Section 4.2. The rest of this section and Section 4.1 explore what may explain this gap.

[17]Median income for the self-employed is $37,500 in the survey data and $18,500 in the administrative data.

Table 1, column (3) examines interest and dividend income, which is included in total money income, but excluded from wage and self-employment, and is available in our survey and administrative sources. Interest and dividends have an important feature that is different than wage + self-employment income – the reporting incentives to the administrative and survey authorities are the same for both wage-earners and self-employed taxpayers because there are third-party reports for all interest and dividend income above $10 in the form of 1099-INT and 1099-DIV filings. So, any differences we observe cannot be due to tax reporting incentives. We examine it only on a significant intensive margin; that is, only among wage-earner and self-employed taxpayers that have positive interest and dividend income and report at least $500 of real interest and dividend income as reported to the survey and administrative authorities.[18] For wage-earners, the gap between survey and administrative sources is remarkably similar across wage + self-employment income and interest + dividend income. However, this is starkly different for the self-employed, who have a much larger gap for wage and self-employment income, but a smaller gap for interest and dividend income. In fact, the gap between their reported interest and dividend income across survey and administrative sources is slightly positive. This rules out one possible mechanism for the difference in the gap between the self-employed and wage-earners: that the self-employed have different survey or administrative record measurement errors that permeate all sources of income.

Two possible mechanisms, both of which focus specifically on self-employment income, remain: (1) the lack of third-party reporting on self-employment income is driving down re-

[18]We choose these restrictions to ensure we look at substantive reporting choices, rather than trivial amounts like $10 or $15, which can lead to large percent differences, despite their relative unimportance. Our conclusions are robust to including the entire intensive interest and dividend margin.

ports to tax authorities relative to what these taxpayers report to survey authorities, and/or (2) measurement error issues are exacerbated for the self-employed in a way that results in substantial overstatement of their earnings in the ASEC. Both may play some role in the overall gap. The rest of this section is focused on robustness checks of the overall gap size examining whether the size of that observed gap is exacerbated by survey mismeasurement. Then Section 4.1 will consider whether the lack of third-party reporting plays a causal role in the size of the gap, controlling for survey mismeasurement and demographic explanations.

Table 1, column (4) reports the gap for total money income instead of wage and self-employment income. This broader income definition is valuable to the degree that we are concerned about miscategorization of income. The self-employed sometimes categorize their self-employment income as wage income (Collins et al., 2019; Abraham et al., 2021), so it is plausible that this type of error could extend to other forms of income as well. For example, a taxpayer could receive dividends, but report it as wages or self-employment income on the ASEC and this could exacerbate the size of the self-employed gap. These estimates are similar to those found in column (1), though the gaps are slightly smaller, suggesting there could be some small merit to this concern. Alternatively, it may be that these additional income types are more similarly reported by both wage and self-employed taxpayers, which is supported by evidence in column (3).

Table 1, column (5) does not aggregate the income data to the taxpayer unit; instead, this regression includes individual-level data on primary respondents only. The rest of the analysis is unchanged. The fact that the estimate in this column is approximately the same as our baseline estimate suggests that our results are not driven by one individual overstating the self-employment income of another member of their household on the ASEC survey by

more than they may have done for wage income.

We consider another survey measurement issue in Table 1, column (6). We exclude those with Schedule E or F income based on an IRS flag for these income sources because the ASEC may not accurately capture income from pass through entities. The gap is slightly larger, but qualitatively similar overall.

Columns (7) and (8) of Table 1 examine mismatch. Household mismatch, where the PVS mistakenly matches an individual's survey record to the wrong individual's administrative record, is another concern that has been previously documented in the literature and could theoretically have a substantial effect on our analysis. Because the self-employed are a relatively small fraction of the population, a mismatched self-employed taxpayer is most likely to be mistakenly matched with a wage-earner. It is plausible to expect that many self-employed taxpayers whose incomes are bunched at the first EITC kink in the administrative data are mismatched with a wage-earner who did not bunch in the survey data. To examine this hypothesis, we restrict our sample in column (7), to only those for whom the Census in most confident in the match – those for whom someone in the taxpaying unit are matched based on their Social Security Number. This sample is limited because the ASEC stopped asking for Social Security Numbers in tax year 2004. In column (8) we also include the group the Census is next most confident in (and most confident after 2004) – those that are matched based on their name, date of birth, and zip code. Our estimates are unaffected by these restrictions.

We consider another type of survey measurement error – rounding – in Figure 4. In this figure, we acknowledge that there are two types of respondents in the survey data – those that do and those that do not round to the nearest $1,000. Mean zero rounding

will not affect the relative size of the self-employed gap (nor will non-mean zero rounding, if done similarly for both self-employed and wage-earners). However, if rounding is non-mean zero and the bias is larger for the self-employed, this would matter. We split the self-employed into groups by whether they round any of their individual income sources to the nearest thousand. The top panel demonstrates that the self-employed who do not round have substantially lower incomes, on average.[19] However, in the bottom panel, we reweight the unround survey and administrative income distributions based on the survey income and demographic distribution of those that do round as described in Section 3. With this reweighting in place, those that do and do not round appear to have similar discrepancies between administrative and survey data. This suggests that rounding does not play a large role in the gap for the self-employed, relative to wage-earners.

In Figure 5, we consider whether the self-employed gap is affected by whether a taxpayer files a Form 1040 with the IRS before or after they respond to the ASEC.[20] Those that file with the IRS first will have recently declared their income and expenses to the IRS, which may allow them to report more accurately to the ASEC. Moreover, this might be particularly helpful for the self-employed who need to think about both revenue and expenses, rather than only their earnings. If this matters, we would expect to see smaller differences between administrative and survey income for those that file with the IRS before completing the ASEC. However, there is another force working in the opposite direction. We know that those that file early are more likely to be those that receive refunds – which makes both tax evasion and lower income levels more likely. We see the latter in the top panel of Figure 5; in

[19] While not shown in the figure, this is also true for wage-earners.

[20] As described in Section 2, we construct a measure that indicates whether the Form 1040 was published in the data bank (which generally corresponds to the filing week) in the first 10 weeks of the year. We use this as a rough proxy for whether the taxpayer filed their 1040 with the IRS before taking the ASEC survey.

the survey data, those that have already filed with the IRS before March are lower income.[21] We reweight so that they share a similar survey income distribution in the bottom panel (as described in Section 3), and this panel suggests that there is more evasion among those that file early, i.e. the bunching at the first EITC kink is larger. This empirically dominates any benefits of knowing more about income before completing the ASEC – the gaps are higher, not lower, for those self-employed that file with the IRS before the ASEC.

While we have not ruled out all forms of survey measurement error, this section has demonstrated that the most obvious possible such sources do not appear to be playing large, substantive roles in driving the size of the self-employed gap relative to that of wage-earners. This supports the idea that while there is evidence in the literature that survey measurement error influences the overall gap, it does not seem to be substantially worse for the self-employed. With this backdrop, we explore another potential mechanism that could explain a sizeable portion of the gap.

4.1 Third Party Reporting as a Possible Mechanism

In this section, we consider whether the tax avoidance or evasion behaviors of the self-employed in the administrative data that do not exist for wage-earners (due to extensive third-party reporting) are a causal mechanism behind the larger gap between survey and administrative sources for the self-employed relative to wage-earners. To do this, we examine the behavior around the first EITC kink, first documented in Figure 1, in more detail.

As with most studies that examine potential tax avoidance or evasion, we cannot directly observe tax avoidance or evasion in the data; instead, we look for statistical "Evidence of

[21] Though not shown here, the same is true among the wage-earner sample.

the Invisible" (Slemrod and Weber, 2012). We know from existing literature that there is substantial bunching of reported incomes at the first EITC kink in tax administrative data for self-employed taxpayers (but little bunching for wage earning taxpayers) (Saez, 2010; Chetty et al., 2013). Based on audit data, the bunching among the self-employed is primarily due to tax evasion (Chetty et al., 2013). Certainly, many other forms of evasion and avoidance occur and are reflected in administrative data; however, the fact that large numbers of taxpayers share a common first EITC kink point gives them a shared reporting target if they wish to maximize their EITC. This allows us to immediately observe this particular form of evasion or avoidance in our figures. An unanswered question – that we examine in this section – is whether this type of documented evasion/avoidance behavior is also reflected in survey responses.[22]

In Section 4, we documented that the peak of the income distribution density for self-employed taxpayers in the administrative data is at the first EITC kink, but not in the survey data (nor in either data set for wage-earners). In Figure 6, we further investigate how much the apparent evasion or avoidance in the administrative data that leads to bunching at the first EITC kink manifests itself in the survey data. We focus in on two groups that have substantial EITC participation and should be similar except for their first EITC kink location – unmarried taxpayers that claim the EITC with one or two dependents (measured as the number of dependents they declare on the EITC). We do not renormalize income and instead plot these two groups separately. We find that they clearly respond to their respective tax avoidance or evasion incentives in the tax administrative data – the peak in

[22]We note that not all individuals at the EITC kink points are engaging in tax evasion, and that in any case we are interested in individual tax evasion only insofar as it is related to survey measurement.

the density for each group aligns with their respective first EITC kink locations and the mean difference in these distributions are more than 50 percent of the difference in their respective first EITC kink locations. In contrast, there is very little difference between their respective survey income distributions. This suggests that the evasion and avoidance that is well-known in the administrative data among the self-employed is not widely reported in the survey data.

We document this more formally in Table 2. Column (1) uses a similar sample as Figure 6 – unmarried EITC claimants with one or two dependents that receive the EITC and have administrative real income less than \$30,000[23] – and interacts SE with an indicator for having two dependents.[24] If a lack of third-party reporting and tax refund incentives nudge tax record income towards the first EITC kink, but have less effect on survey income, we expect the gap between survey and administrative records to be smaller (in absolute value) for two-dependent taxpayers because the first EITC kink is at a higher income level for these taxpayers; the gap may even become positive to the extent that taxpayer's increase, rather than decrease, their income to reach the first EITC kink. This is exactly what we find: the coefficient on the interaction term suggests that, relative to wage-earners, the gap is approximately 21 percentage points smaller for two dependent relative to one dependent self-employed taxpayers. In fact, the cumulative self-employment gap for two dependent taxpayers is approximately zero, and we can see this in Figure 6.[25] In Figure 6, the bunched

[23]The \$30,000 administrative income restriction is in place so that all taxpayers in this sample are in the EITC range, and hence are eligible for the EITC (and if they choose to file we will observe the number of EITC dependents). The results in this table are highly robust to an alternative specification in which we measure dependents based on the number of 1040 tax return dependents and focus on those with below-median ASEC income.

[24]For all interacted variables in this table, the non-interacted variables are included in the regression as well, but they are generally not reported as they are not the focus of this analysis.

[25]Also note that the gap among this income group is the largest for those excluded from this regression –

mass in administrative income at the first EITC kink appears to be coming in roughly equal amounts from both sides for those with two dependents (relative to the survey distribution). In contrast, for those with one dependent where the first EITC kink is lower, more of the bunched mass appears to be coming from above the first EITC kink.

In column (2), we include year fixed effects and interactions between SE and \$1,000 survey income bins. This leaves the estimates roughly unchanged. Column (3) includes married taxpayers and an interaction between marital status and self-employment and we again find similar results on the interaction with two dependents. In column (4), we control for indicators of the survey measurement concerns we examined in section 4 and some demographic variables as well as their interactions with SE. The demographic variables include measures of gender, marital status, age, education, full-time work, and home-ownership. Because we did not find clear evidence that survey measurement concerns significantly influenced the gap above, it is unlikely that they significantly change our estimates on two relative to one dependent here, but we confirm for completeness. Our expectations are confirmed and the coefficients on these survey measurement errors interacted with SE align with the conclusions we drew from the figures and regressions regarding those measurement errors in the previous section.[26]

Overall, Figure 6 and Table 2 provide strong supporting evidence that an important causal mechanism for the gap between survey and administrative data in self-employed tax-

those with no dependents – and this is not surprising since their first EITC kink is even lower than for those with dependents.

[26]Though not reported here, we also consider additional measurement error issues in the overlap sample, where we observe all taxpayers in two consecutive years. The gap for the self-employed does not change in the second year of response relative to the first, nor does the gap change for an especially good or bad survey year (>10 percent change relative to the other year). And neither of these additional survey measurement control variables influence the conclusions of Table 2.

payers is tax evasion or avoidance opportunities engaged in by this group that are not available to wage-earners.

4.2 Implications

Although understanding survey measurement is important, the existence of differences between survey and administrative self-employment income (and the sources of these differences therein) could be merely an academic exercise if they do not have impact beyond the survey in question. However, there is reason to believe that there may be downstream measurement implications given our findings, which could potentially change important inferences about the state of the economy or social characteristics of the United States. In this section, we consider a few simple examples of such implications: how does the self-employed poverty rate differ between survey and administrative sources, are the rank orderings of taxpayers by income preserved or shifted across the two sources for self-employed taxpayers, and when we examine the transition into self-employment, are the conclusions we draw affected by whether we use survey or administrative data?

Given we found self-employed earnings are much lower in the administrative data, relative to the survey data, we expect that the measured poverty rates among the self-employed will differ too. We calculate the poverty rate based on everyone that reports positive survey and administrative wage + self-employment income. While we recognize that this does not capture all poverty, we select this measure to keep our focus on the intensive reporting margin that we have been considering in this paper. The self-employed poverty rates are almost double in the administrative data (45 percent) relative to the survey data (25 percent).

Poverty differences between survey and administrative sources are much smaller for the overall population or wage-earners (Hokayem et al., 2015).

It remains uncertain whether the self-employed are just shifted to the left in the administrative data or whether there is a large reshuffling of the rank-order of taxpayers across the two data sets. The latter is likely if evasion occurs in the administrative data, and occurs to differing degrees across taxpayers. We examine this in Table 3. We find that, among wage-earners, the likelihood of staying in the same quintile in both data sets is 62% on average, with the highest likelihoods in the bottom and top quintiles. In contrast, it is much lower – 39% – for the self-employed, and the declines in persistence occur across all quintiles.

As workers transition into self-employment, they see a decline in income in the first year of their transition, but the magnitude is about twice as large in the administrative data – 13.5 percent (survey) compared to 27 percent (administrative). This discrepancy is statistically significant (p-value = 0.015) and not surprising given the findings above. But it highlights a novel phenomenon – the same person makes different reporting decisions for survey relative to administrative sources once they transition to self-employment.

5 Conclusion

While the outcomes of self-employed individuals in survey or administrative data have been studied (e.g. Bruce, 2000; Carroll et al., 2000; Gentry and Hubbard, 2000; Gale and Brown, 2013), we know relatively little about how and why these two sets of income reports compare, because there has not been a comprehensive intensive margin study of survey income measurement among this sub-population. In this study, we abstract away from previously doc-

umented miscategorization issues (e.g. self-employed individuals declaring that they make wage income and vice versa), and compare the wages + self-employment income of the self-employed relative to their wage-earning counterparts in linked survey and administrative data. Our analysis is focused on taxpayers that report positive amounts of income to both survey and administrative sources. We find dramatically larger gaps for the self-employed; they report about 40 percent more to the survey than they report to the IRS, on average, netting out any baseline gaps we observe among wage-earners. This result is highly robust and does not appear to be substantially driven by survey measurement issues. Moreover, the self-employed do not appear to have a different survey income reporting propensity across all types of income as their interest and dividend income reports are fairly close across both survey and administrative sources.

Identifying a potential mechanism behind this larger gap among the self-employed is a particularly challenging problem, as it is quite plausible that both the survey and administrative data sources are biased in some ways. Nevertheless, it's an essential step toward a better understanding of survey measurement quality for the self-employed. We hope this increased understanding will lead to survey measurement improvements for this sub-population going forward. Administrative data, by construction, reflects the tax avoidance and evasion choices of all taxpayers. It is well known that, because of the structure of the tax code and lack of informational reporting and withholding for the self-employed, these taxpayers more frequently engage in tax avoidance and evasion activities. To the extent that these choices are not also reflected in survey earnings reports, this provides a plausible mechanism for the larger difference between these two sources among self-employed taxpayers. By exploiting features of the first EITC kink, a well-known focal point of tax evasion in the administrative

tax data, we find that much of the evasion/avoidance activity in the administrative data does not appear to be present in the survey data.

The fact that the mean incomes of self-employed individuals in survey and administrative data differ so greatly indicates that the choice of income data source will matter for most analyses of self-employed taxpayers and the determination of their poverty rates. We highlighted a few of these implications in Section 4.2. For example, the intensive margin poverty rate for self-employed and the income loss in the year of transition to self-employment are both about twice as high in the administrative data. This research highlights areas for future inquiry and efforts to improve income measurement, as in Bee et al. (2023). Unlike wage income, whose extensive and overlapping third party reporting means that arguably administrative records can be seen as a ground truth, self-employment income's lack of third party reporting may point towards the use of survey responses in some circumstances where administrative records may reflect misrepresentations to tax authorities.

References

Abowd, John M. and Martha H. Stinson, "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *The Review of Economics and Statistics*, 2013, *95* (5), 1451–1467.

Abraham, Katharine G., John C. Haltiwanger, Claire Hou, Kristin Sandusky, and James R. Spletzer, "Reconciling Survey and Administrative Measures of Self-Employment," *Journal of Labor Economics*, 2021, *39* (4), 825–860.

Ballard, Charles and Sanjay Gupta, "Perceptions and Realities of Average Tax Rates in the Federal Income Tax: Evidence from Michigan," *National Tax Journal*, 2018, *71* (2), 263–294.

Bee, Adam and Jonathan Rothbaum, "The Administrative Income Statistics (AIS) Project: Research on the Use of Administrative Records to Improve Income and Resource Estimates," SEHSD Working Paper Number 2019-36 2019.

_ , Joshua Mitchell, Nikolas Mittag, Jonathan Rothbaum, Carl Sanders, Lawrence Schmidt, and Matthew Unrath, "National Experimental Wellbeing Statistics - Version 1," Working Papers 23-04, Center for Economic Studies, U.S. Census Bureau February 2023.

_ , Sandy Clark, Kevin Rinz, Jonathan Rothbaum, Chase Sawyer, John Voorheis, Thomas Hertz, Pat Langetieg, and Mark Payne, "Improving Income Measurement in the American Community Survey Using Expanded 6103(n) Income Data from the Internal Revenue Service," Mimeo 2021.

Bollinger, Chris R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak, "The Good, The Bad and The Ugly: Measurement Error, Non-response and Administrative Mismatch in the CPS," Working Paper 2018.

Bollinger, Christopher R., "Measurement Error in the Current Population Survey: A Nonparametric Look," *Journal of Labor Economics*, 1998, *16* (3), 576–594.

_ , Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak, "Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch," *Journal of Political Economy*, 2019, *127* (5), 2143–2185.

Bound, John and Alan B. Krueger, "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?," *Journal of Labor Economics*, 1991, *9* (1), 1–24.

_ , Charles Brown, and Nancy Mathiowetz, "Chapter 59 - Measurement Error in Survey Data," in James J. Heckman and Edward Leamer, eds., *James J. Heckman and Edward Leamer, eds.*, Vol. 5 of *Handbook of Econometrics*, Elsevier, 2001, pp. 3705–3843.

Bruce, Donald, "Effects of the United States Tax System on Transitions into Self-Employment," *Labour Economics*, 2000, *7* (5), 545–574.

Carroll, Robert, Douglas Holtz-Eakin, Mark Rider, and Harvey S. Rosen, "Income Taxes and Entrepreneurs' Use of Labor," *Journal of Labor Economics*, 2000, *18* (2), 324–351.

Chetty, Raj, John N. Friedman, and Emmanuel Saez, "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review*, December 2013, *103* (7), 2683–2721.

Collins, Brett, Andrew Garin, Emilie Jackson, Dmitri Koustas, and Mark Payne, "Is Gig Work Replacing Traditional Employment? Evidence from Two Decades of Tax Returns," *Working Paper*, 2019.

DeBacker, Jason, Bradley Heim, Anh Tran, and Alexander Yuskavage, "Tax Noncompliance and Measures of Income Inequality," *Tax Notes Federal*, 2020.

Gale, William and Samuel Brown, "Small Business, Innovation, and Tax Policy: A Review," *National Tax Journal*, 2013, *66* (4), 871–892.

Gentry, William M. and R. Glenn Hubbard, "Tax Policy and Entrepreneurial Entry," *American Economic Review*, May 2000, *90* (2), 283–287.

Gideon, Michael, "Do Individuals Perceive Income Tax Rates Correctly?," *Public Finance Review*, 2017, *45* (1), 97–117. PMID: 29238156.

Hokayem, Charles, Christopher Bollinger, and James P. Ziliak, "The Role of CPS Nonresponse in the Measurement of Poverty," *Journal of the American Statistical Association*, 2015, *110* (511), 935–945.

Hurst, Erik, Geng Li, and Benjamin Pugsley, "Are Household Surveys Like Tax Forms? Evidence from Income Underreporting of the Self-Employed," *The Review of Economics and Statistics*, 03 2014, *96* (1), 19–33.

Kapteyn, Arie and Jelmer Y. Ypma, "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data," *Journal of Labor Economics*, 2007, *25* (3), 513–551.

Layne, Mary, Deborah Wagner, and Cynthia Rothhaas, "Estimating Record Linkage False Match Rate for the Person Identification Validation System," CARRA Working Papers 2014-02, Center for Economic Studies, U.S. Census Bureau July 2014.

Meijer, Erik, Susann Rohwedder, and Tom Wansbeek, "Measurement Error in Earnings Data: Using a Mixture Model Approach to Combine Survey and Register Data," *Journal of Business and Economic Statistics*, 2012, *30*, 191–201.

Pischke, Jörn-Steffen, "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study," *Journal of Business Economic Statistics*, 1995, *13* (3), 305–314.

Saez, Emmanuel, "Do Taxpayers Bunch at Kink Points?," *American Economic Journal: Economic Policy*, August 2010, *2* (3), 180–212.

Slemrod, Joel and Caroline Weber, "Evidence of the invisible: Toward a credibility revolution in the empirical analysis of tax evasion and the informal economy," *International Tax and Public Finance*, 2012, *19* (1), 25–53.

_ **and Jon Bakija**, *Taxing Ourselves*, 4th ed., Cambridge, Massachusetts: The MIT Press, 2008.

Tables

Table 1: Gap Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Self-Employment (SE)	-0.399^{***}	-0.486^{***}	0.200^{***}	-0.383^{***}	-0.423^{***}	-0.435^{***}	-0.352^{***}	-0.397^{***}
	(0.008)	(0.007)	(0.033)	(0.015)	(0.008)	(0.009)	(0.022)	(0.008)
Constant	-0.082^{***}	-	-0.079^{***}	-0.071^{***}	-0.049^{***}	-0.071^{***}	-0.092^{***}	-0.084^{***}
	(0.001)	-	(0.008)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)
Observations	659,000	659,000	27,900	580,000	495,000	567,000	105,000	650,000
Income measure	Wage+SE	Wage+SE	Int+Div	TMI	Wage+SE	Wage+SE	Wage+SE	Wage+SE
Includes Year FE & CPS 1K Bins?	No	Yes	No	No	No	No	No	No
Tax-Unit (TU) or Individual Primary Respondent (IPR)?	TU	TU	TU	TU	IPR	TU	TU	TU
Restriction?	-	-	-	-	-	No SchedE/F	PVS V	PVS S&V

Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: Standard errors are in parentheses. Standard errors are clustered by taxpayer. *5% significance level. **1% significance level. ***0.1% significance level. The SE coefficients are estimated from equation (1); in this table, no additional X_{it} are included unless otherwise noted. Unless specified otherwise, the dependent variable in this regression is the percent gap in administrative - survey wage + self-employment earnings. Column (2) adds year fixed effects and $1,000 survey earnings bins. Column (3) constructs the dependent variable with interest and dividend income and restricts to taxpayers that have at least $500 of this income reported to both survey and administrative authorities. Column (4) replaces the income measure in the dependent variable with total money income. Column (5) returns the dependent variable measure to wage + self-employment income but uses individual data on primary respondents instead of data aggregated to the taxpayer unit. Column (6) excludes taxpayers with Schedule E or F IRS income flags. Column (7) includes on those with PVS code "V", that is only those with an SSN match between their ASEC and DER records (these are only available in survey years 2001-2004). Column (8) is the same as (7), but also includes the next best type of match, PVS code "S", which are matches on name, date of birth, and zip code.

Table 2: Gap Estimates by Number of Dependents

	(1)	(2)	(3)	(4)
SE	-0.254^{***}	-	-	-
	(0.023)			
SE x 2 EITC Dependents	0.214^{***}	0.191^{***}	0.166^{***}	0.161^{***}
	(0.034)	(0.018)	(0.015)	(0.015)
SE x Not Round				-0.008
				(0.018)
SE x IRS before ASEC				-0.163^{***}
				(0.015)
SE x PVS S&V				-0.016
				(0.053)
Observations	49,000	49,000	66,500	55,000
ASEC 1K bins x SE	No	Yes	Yes	Yes
Married Households Included	No	No	Yes	Yes
Demographics x SE included	No	No	No	Yes

Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: Standard errors are in parentheses. Standard errors are clustered by taxpayer. *5% significance level. **1% significance level. ***0.1% significance level. The dependent variable in this regression is the percent gap in administrative - survey wage + self-employment earnings. We estimated (1) and these are select coefficients from that regression. Column (1) includes all unmarried households with one or two EITC dependents with real 2010 administrative income $<$ \$30,000. Except for the constant, all estimated coefficients are reported. Column (2) adds interactions between SE and survey \$1,000 income bins and year fixed effects. Because of the interactions, the SE coefficient on it's own is no longer informative, so we suppress it here. Column (3) adds married taxpayers and includes an indicator for marital status along with an interaction between SE and marital status (these are suppressed in this output as they are control variables that are not of direct interest). Column (4) adds additional covariates. We add several measures of survey accuracy and interact them with SE; we report the interactions in the table. This column also adds a number of demographics and interacts them with SE. These are suppressed as they are not of primary interest. The demographics included are: gender, marital status, age, education, full-time work, and homeownership.

Table 3: Quintile Transition Matrix between Survey and Administrative Data

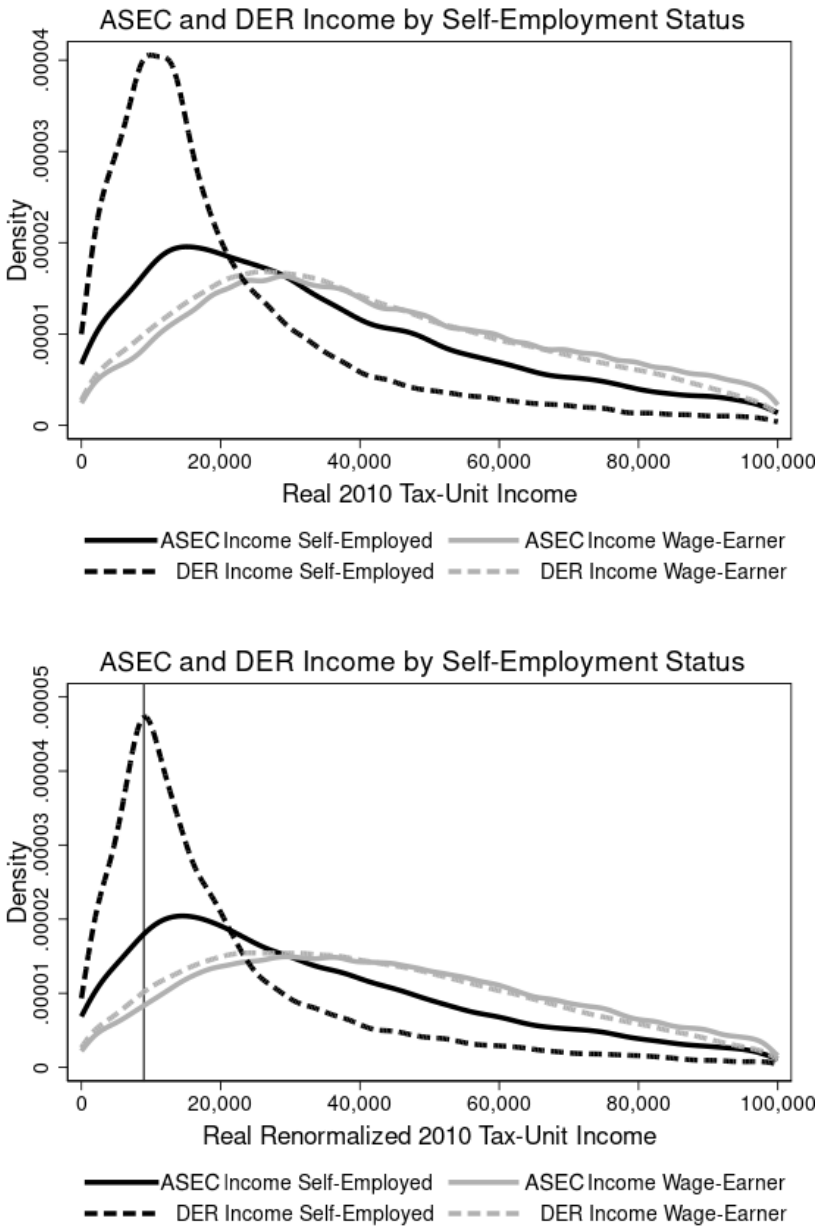
	Wage-Earners				
DER \ ASEC	Bottom	Second	Third	Fourth	Top
Bottom	0.703	0.168	0.071	0.036	0.022
Second	0.215	0.554	0.143	0.059	0.029
Third	0.055	0.211	0.537	0.139	0.057
Fourth	0.021	0.049	0.204	0.583	0.144
Top	0.009	0.017	0.044	0.184	0.747

	Self-Employed				
DER \ ASEC	Bottom	Second	Third	Fourth	Top
Bottom	0.409	0.227	0.167	0.129	0.068
Second	0.361	0.301	0.180	0.113	0.045
Third	0.145	0.336	0.260	0.183	0.076
Fourth	0.061	0.121	0.318	0.333	0.167
Top	0.023	0.030	0.075	0.241	0.632

This table is a quintile transition matrix between administrative (DER) and survey (ASEC) earnings separately for self-employed and wage-earner households. There are 30,500 observations contributing to the self-employed transition matrix and 501,500 observations contributing to the wage-earner transition matrix. This transition matrix pools data from tax years 2000-2015.

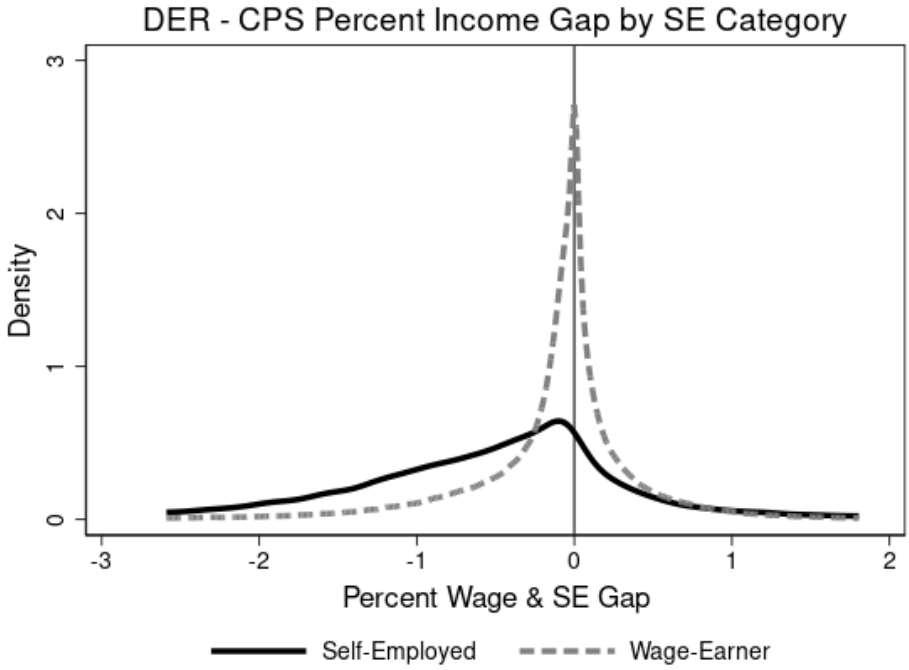
Figures

Figure 1: Kernel Density Plots by Self-Employment Status



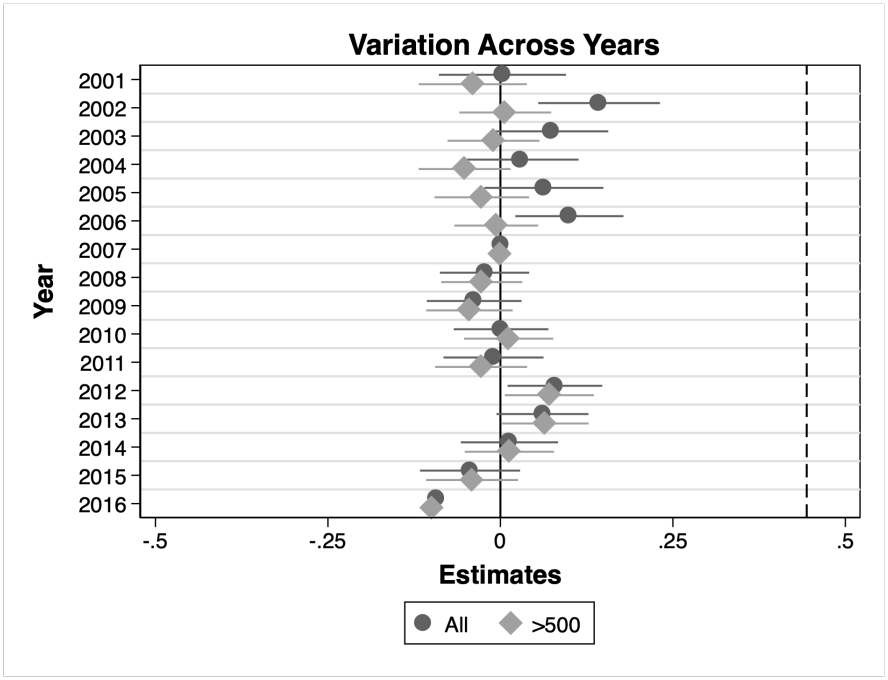
Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: The top figure plots a Gaussian kernel of positive survey (ASEC, solid lines) and administrative (DER, dashed lines) income for all taxpayers whose survey and administrative income is less than $100,000 separately for the self-employed (black lines) and wage-earners (gray lines). There are 509,500 tax-unit observations in this figure and 5.2 percent of them are self-employed. Average survey earnings for self-employed taxpayers is $34,500, relative to their administrative earnings of $21,500. In contrast, wage earnings are relatively close – $44,000 in the survey and $41,500 in the administrative data. The bottom panel repeats the top panel with income renormalized so that all taxpayers have the same first EITC kink point as described in Section 3. There is a vertical line at the first EITC kink. All kernel densities use a Gaussian kernel because of U.S. Census Bureau disclosure rules.

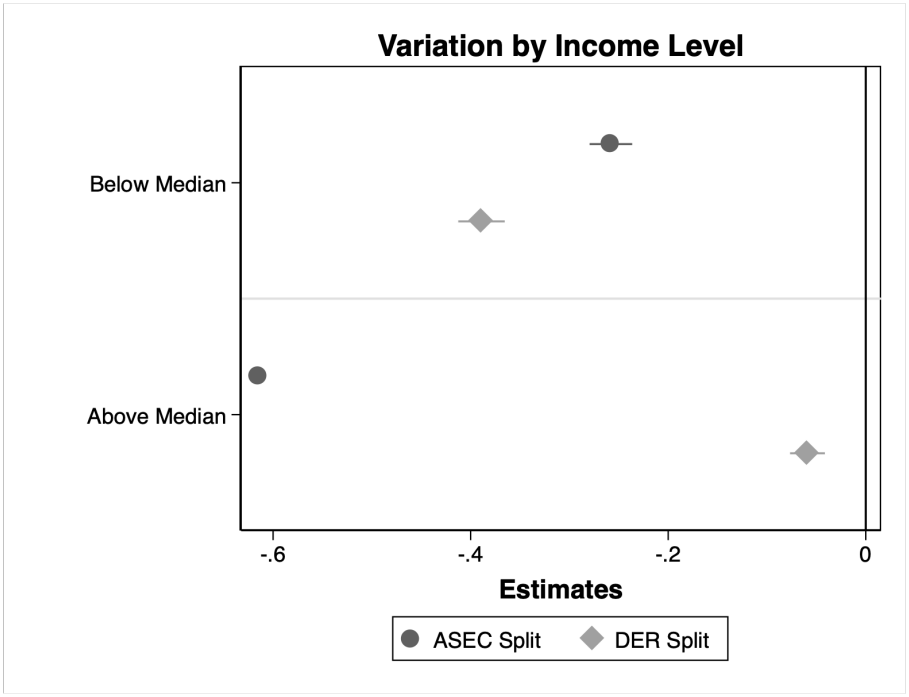
Figure 2: Kernel Density Plots by Self-Employment Status



Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: This figure subtracts the log of taxpayer's administrative income from the log of their survey income (so a negative gap implies that their administrative income is lower than their survey income) and plots a kernel density of this difference for the self-employed (black line) and wage-earners (gray dashed line) for the same tax units that are included in the top panel of Figure 1. The average gap for the self-employed is -50.0%; the average gap for wage-earners is -6.8%. All kernel densities use a Gaussian kernel because of U.S. Census Bureau disclosure rules.

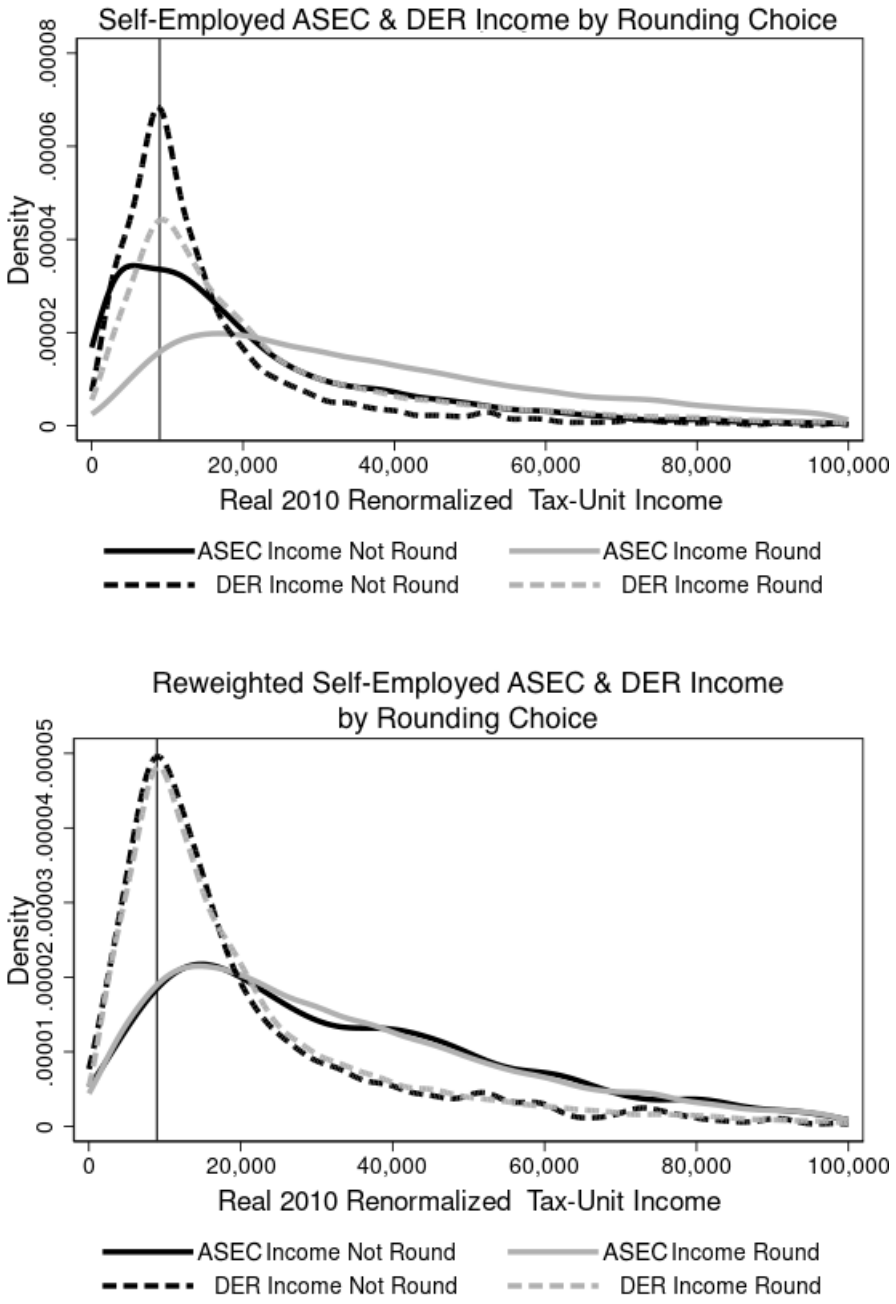
Figure 3: Variation by Year and Income Levels





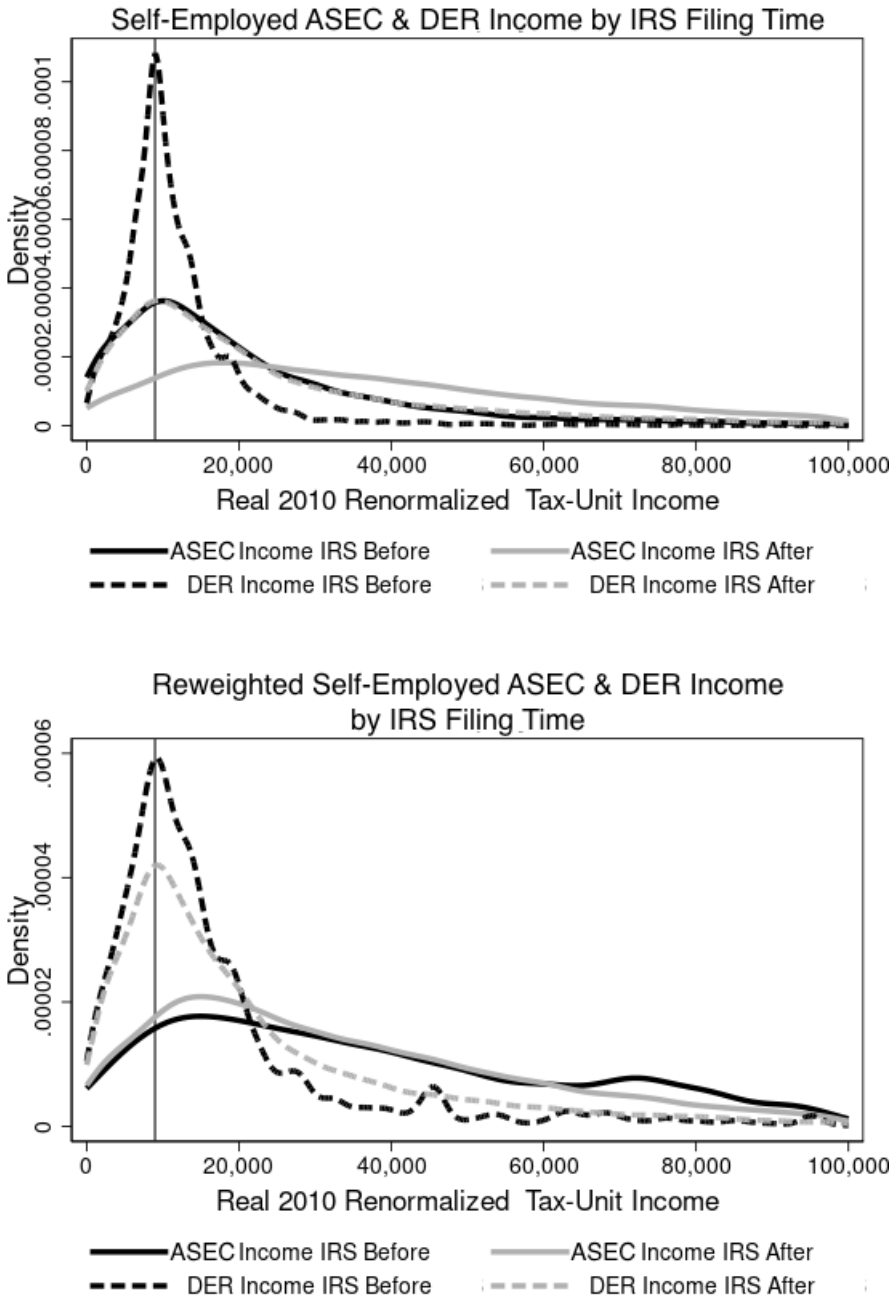
Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: The top panel interacts SE in equation (1) with year indicators and plots the estimated self-employed gap (relative to the wage-earner gap) separately for each year. 2007 is the reference year. We overlay a second set of estimates that excludes those with survey income less than $500 as very small amounts of income can generate extreme outliers. The vertical dashed line is the absolute value of the main estimate; it highlights the variation across years is small relative to the baseline estimate. The bottom panel interacts SE in equation (1) with an indicator for being above median survey income. We separately repeat this exercise using an indicator based on median administrative income. Median income for the self-employed in the ASEC is $37,500 and in the DER it is $18,500. For each, we report the total gap (the coefficient on SE for those below the median and the sum of the coefficients on SE and $SExAboveMedian$ for those above the median) for the self-employed (relative to wage-earners).

Figure 4: Kernel Density Plots by Rounding Status



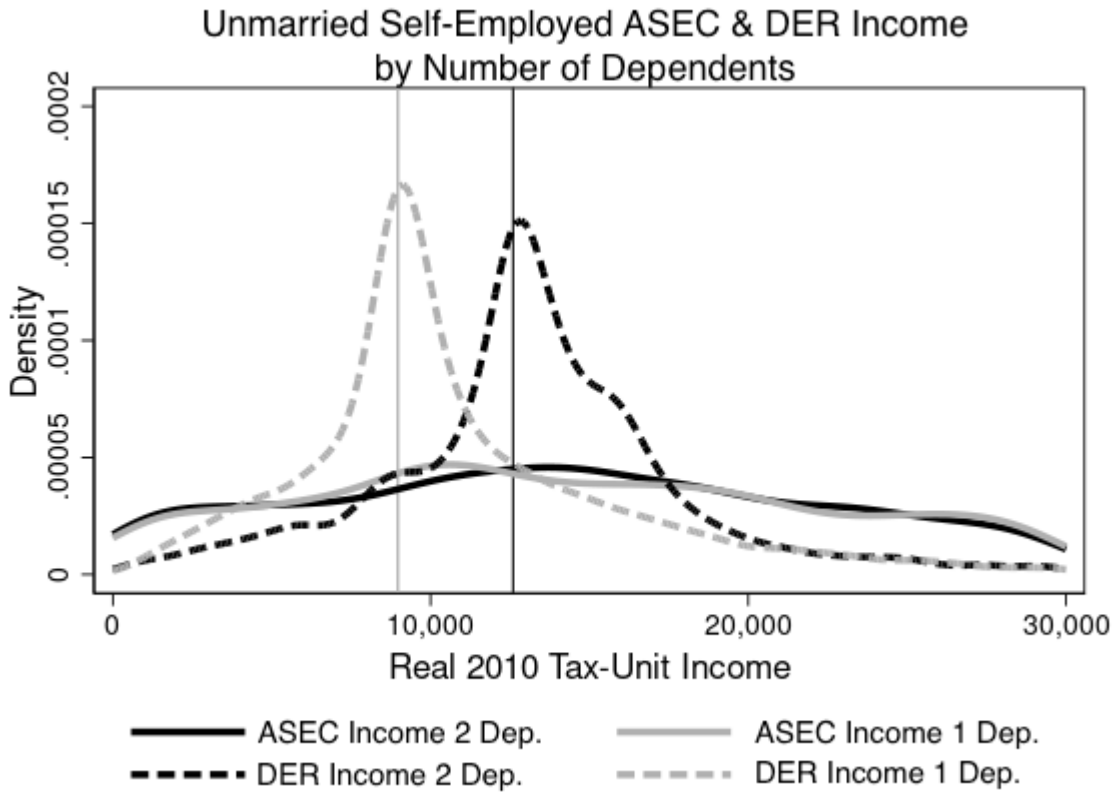
Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: The top panel replicates the bottom panel of the self-employed density plots in 1 separately for taxpayers that round (13%) and don't round (87%) to the nearest $1,000. These figures restricts income to be at least $1,000 so that both densities begin at the same place. There are 26,000 tax-unit observations in these figures. The bottom panel reweights the unround survey and administrative income distributions based on the survey income and demographic distribution of those that do round. All kernel densities use a Gaussian kernel because of U.S. Census Bureau disclosure rules.

Figure 5: Kernel Density Plots by IRS Filing Time



Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: The top panel replicates the bottom panel of the self-employed density plots in 1 separately for taxpayers that file with the IRS before they interview with the ASEC (19%) and for those that file during or after they interview with the ASEC (81%). There are 26,500 tax-unit observations in these figures. The bottom panel reweights the survey and administrative income distributions of those that file their tax return before the ASEC based on the survey income and demographic distribution of those that filing during or after their ASEC interview. All kernel densities use a Gaussian kernel because of U.S. Census Bureau disclosure rules.

Figure 6: First EITC Kink Kernel Density Plot by Number of Dependents



Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: This figure repeats the top panel of Figure 1. It plots survey (ASEC) and administrative (DER) self-employment income separately for unmarried one- (45%) and unmarried two-dependent (55%) taxpayers. The number of dependents is determined by the number of EITC dependents they claim (so this figure only examines those that claim the EITC). There are 3,500 tax-unit observations in this figure. The gap between the survey income distributions of these two groups is only $500. However, the gap in administrative income is $2,000. This average difference in administrative income is more than 50% of the difference in first EITC kink locations. The first EITC kink locations are marked with vertical black lines (in 2010, the EITC kink for one dependent is $8,970 and for two dependents it is $12,590). All kernel densities use a Gaussian kernel because of U.S. Census Bureau disclosure rules.

Appendices

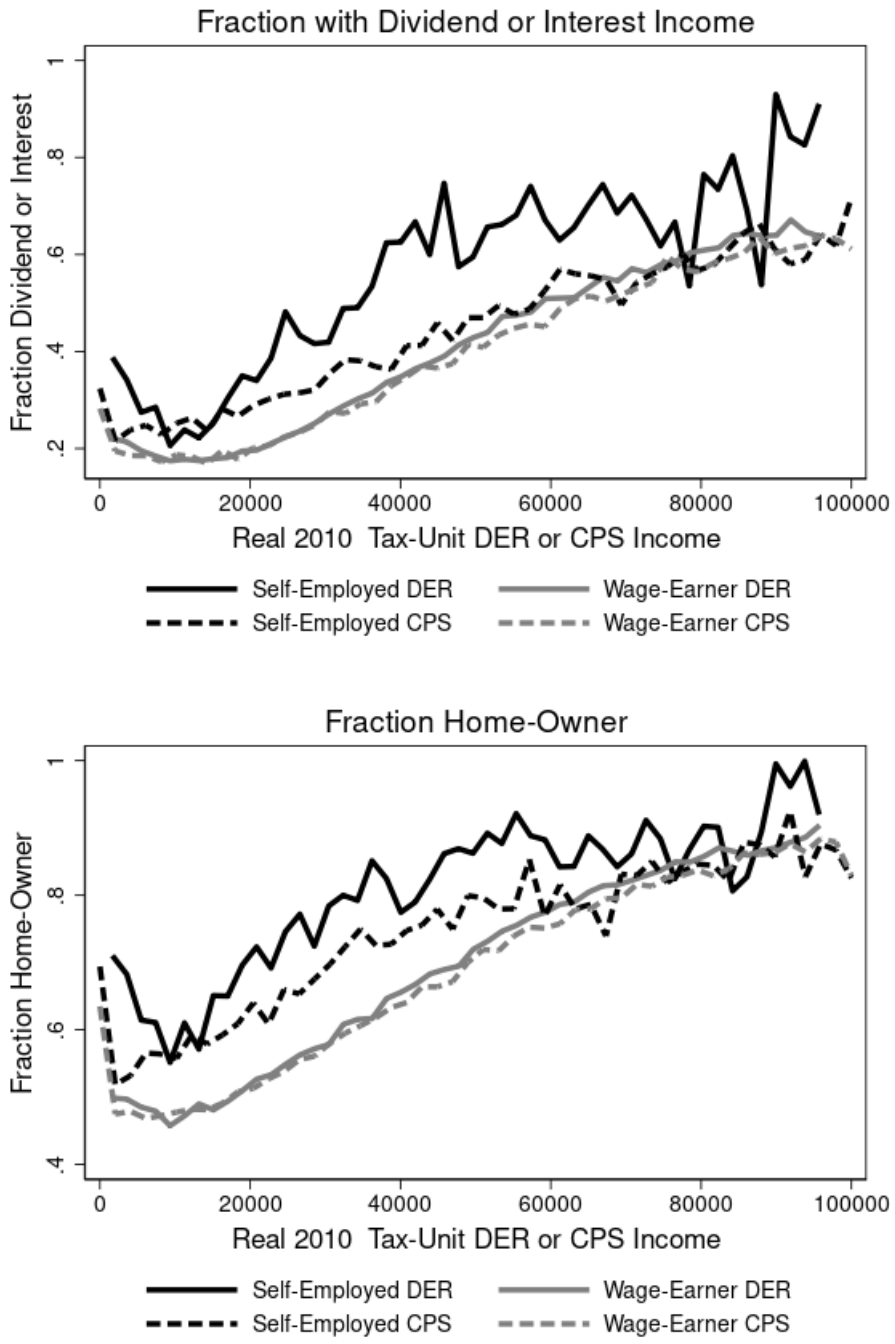
A Income Measures and Wealth Proxies

Hurst et al. (2014) assume wage-earners and the self-employed have the same marginal propensity to consume out of their income. When they observe a much higher marginal propensity to consume for the self-employed, they argue that this is because they are under-reporting their income to survey authorities, just as they do to tax authorities. We don't have good overall consumption measures in the ASEC, so we don't replicate that extensive effort here. However, we do have a few measures of wealth for which the same intuition should apply – we expect our measures to increase with income and the likelihood of having an item correlated with wealth should not vary by self-employed status at a given income level if self-employed income is accurately reported to the survey. If, on the other hand, self-employed income is under-reported to the survey, then the likelihood of having a given form of wealth will be higher for the self-employed at a given level of reported survey income relative to wage-earners. Our two measures of wealth are whether the taxpayer has interest or dividend income according to the administrative records, and whether the taxpayer owns a home according to the survey. We see a pattern consistent with Hurst et al. (2014), at least up to about $60,000, in Figure A.1; in the survey data (dashed lines), the likelihood that the taxpayer has each wealth proxy is higher, at a given survey income level, for the self-employed than wage-earners.[27] If we plot these wealth proxies across administrative data (solid lines), the plot is very similar to the survey income plot for wage-earners. But the likelihood of having each of these wealth proxies are substantially higher for the self-employed at a given administrative income level relative to any given survey income level. And this generally persists across the entire income distribution that we are able to plot.

[27]Note, there is an alternative hypothesis to the Hirst, Li, and Pugsley explanation: self-employed taxpayers have higher wealth for a given level of income and consume more out of that higher wealth. If that were the case, then the surveys may not actually under-report self-employed income. However, that does not change the additional difference we document in the administrative data.

This provides yet another piece of evidence of more evasion/avoidance in the administrative data than the survey data for self-employed taxpayers.

Figure A.1: Relation Between Income Measures and Wealth Proxies



Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: The top panel plots the likelihood the taxpayer has interest or dividend income according to administrative records across the survey (ASEC) and administrative (DER) income distributions up to $100,000. The bottom panel repeats this exercise for the whether the taxpayer owns a home according to the survey. There are 499,500 tax-unit observations in each of these figures.

B Appendix Tables

Table B.1: Gap Estimates by Self-Employment Measure

	(1)	(2)	(3)	(4)
SE	-0.344^{***}	-0.399^{***}	-0.509^{***}	-0.239^{***}
	(0.007)	(0.008)	(0.009)	(0.011)
Constant	-0.080^{***}	-0.082^{***}	-0.082^{***}	-0.094^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
Observations	659,000	659,000	659,000	659,000
SE Measure	50% DER	75% DER	100% DER	75% CPS

Source: ASEC, SSA DER and IRS 1040s, tax years 2000-2015. Notes: Standard errors are in parentheses. Standard errors are clustered by taxpayer. *5% significance level. **1% significance level. ***0.1% significance level. The dependent variable in this regression is the percent gap in administrative - survey earnings. Column (1) defines SE as any taxpayer for whom at least 50 percent of their administrative income is self-employment income. Column (2) repeats Column (1) using a 75 percent threshold (this is our baseline measure used in other tables throughout the paper). Column (3) repeats Column (1) using a 100 percent threshold. Column (4) repeats (2) but defines SE based on survey rather than administrative income.